

Яндекс

1. Как устроен Поиск по блогам
2. Блогосфера Беларуси 2010

Антон Волнухин
Минск, 15 апреля 2010

Яндекс

Как устроен Поиск по блогам

Антон Волнухин
Минск, 15 апреля 2010

Что это такое?

Что такое Яндекс.Поиск по блогам и зачем он нужен?

На каких принципах основан?

Что такое Поиск по блогам?

Поиск по мнениям. Общественное мнение в интернете

- Поиск по текстам, где люди говорят от первого лица, возможность сравнить обсуждаемость чего-либо:
 - что другие говорят о вас или ваших действиях
 - что пишут о товаре, который вы собираетесь купить
 - что пишут о вашей компании
 - что пишут о каком-то событии
- Наиболее обсуждаемые темы и самое популярное в интернете сегодня

Например, Алиса в Стране чудес

Найти

расширенный поиск



Главные темы дня

- ▶ [Новый логотип Femen от Артемия Лебедева](#)
- ▶ [В редакции журнала The New Times изымают документы](#)
- ▶ [Apple одобрила браузер Opera Mini для iPhone](#)

За последние три дня 202 записи посвящено трём самым популярным сегодня темам.

Остальные темы

- [Акция Федерации автовладельцев России против «мигалок»](#)
- [Саммит по ядерной безопасности в США](#)
- [Самолет президента Польши разбился при посадке](#)
- [Обращение Алексея Дымовского к президенту Медведеву](#)

[Формы поиска для вашего сообщества](#)

 Из каталога: [Юмор](#) 61 блог [Творчество](#) 280 [Развлечения](#) 318 [Дом](#) 171 [Технологии](#) 327 [Деловые](#) 171 [Ещё...](#)

Самое популярное и обсуждаемое в интернете

Сервисы

	LiveJournal	72 676
	LiveInternet	23 817
	Блоги@Mail.Ru	21 970
	Я.ру	20 762
	Diary.ru	13 683
	Blogger.com	5 339
	Love Planet	4 611
	BabyBlog.ru	3 841
	24open.ru	2 851
	Дневники на MyLove.ru	2 058

Всего 106 сервисов

Блоги

	bigdan	227 645
	dnugoi	178 487
	tema	165 421
	НОВОСТИ В ФОТОГРАФИИ	158 347
	Интернет-журнал ETODA	131 938
	teh_nomad	128 862
	Людмила Каганов: дневник	118 351
	Lifehacker.ru	114 691
	uborshizza	111 478
	Блог Яндекса	103 114

Всего 17 901 903 блога

Запросы

- [Артур Гурский](#)
- [Права](#)
- [Владислав Сижорский](#)
- [Общество Сених Ведейков пилсудский](#)
- [Грибы](#)
- [Валентин Валентинов](#)
- [Артемия Лебедева](#)
- [Вайда Анджей](#)
- [статьи Рабы ОМОНа](#)

Топ 50 запросов

Обсуждаемые новости

- [Поляки благодарят россиян за поддержку в дни национальной трагедии](#)
193 мнений
- [Министр культуры обвинил россиян в исчезновении буха алфавита](#)
189 мнений
- [Opera Mini для iPhone скачали более миллиона раз](#)
187 мнений
- [В Пентагоне считают, что Иран за год может произвести достаточное для создания ядерного боезаряда количество урана](#)
175 мнений
- [После землетрясения в Китае зафиксированы 774 повторных толчка](#)
169 мнений

[все обсуждаемые новости](#)

Смотрите [рейтинги записей](#), основанные на API Поиска по блогам и [описание того](#), как создавать на базе API свои рейтинги.

Сервисы микроблогов

Фильмы



Почему социальный поиск важен?

свежесть

важный источник информации о важном для людей

возможность изучать живое общение

иногда людям нужен ответ от другого живого
человека, оценочные суждения

структурированность

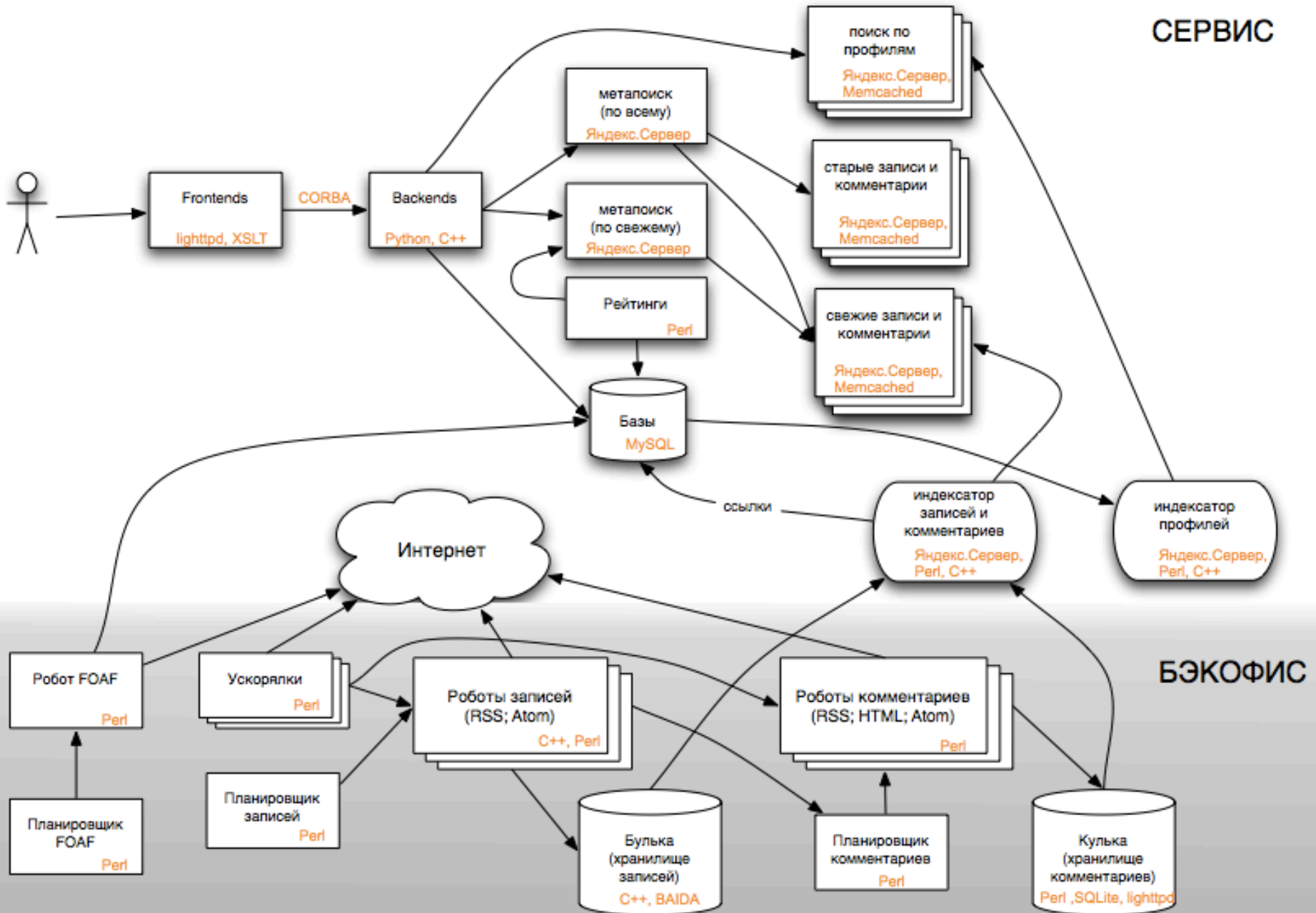
большой объём

Масштабы

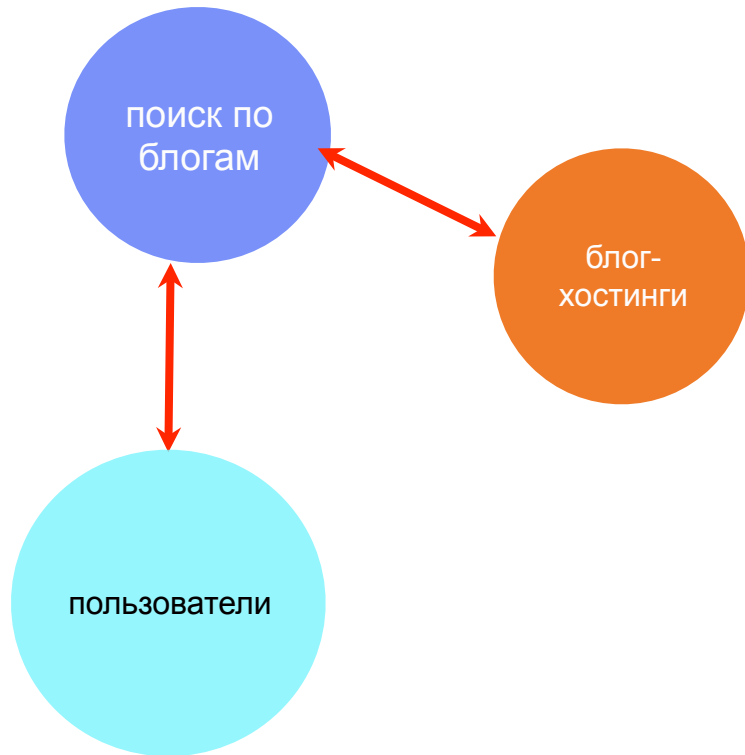
- Более **миллиона** записей и комментариев из блогов и форумов каждый день
- Почти **25 миллионов** источников
- Всего около **полутора миллиардов** документов

Поиск по блогам – это почти **одна пятая** от поиска по всему русскоязычному интернету по количеству элементов индексации

Внутреннее устройство

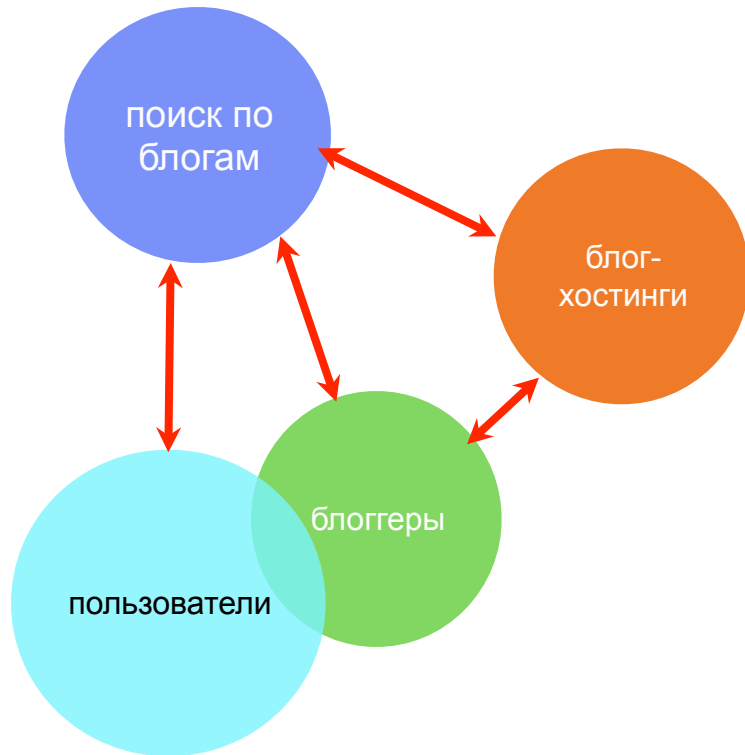


Модель сервиса



- партнёрство и взаимодействие между участниками:
 - блогхостинги
 - пользователи
- быть зеркалом блогосферы
- полностью автоматический сервис
- единые правила для партнёров
- открытые форматы (RSS, ATOM, FOAF)
- все наши API доступны публично

Модель сервиса



- партнёрство и взаимодействие между участниками:
 - блоггеры
 - блогхостинги
 - пользователи
- быть зеркалом блогосферы
- полностью автоматический сервис
- единые правила для партнёров
- открытые форматы (RSS, ATOM, FOAF)
- все наши API доступны публично

Содержание

Что это и зачем

1. Поиск
2. Пульс блогосферы
3. Рейтинги
4. Темы дня
5. Открытые данные

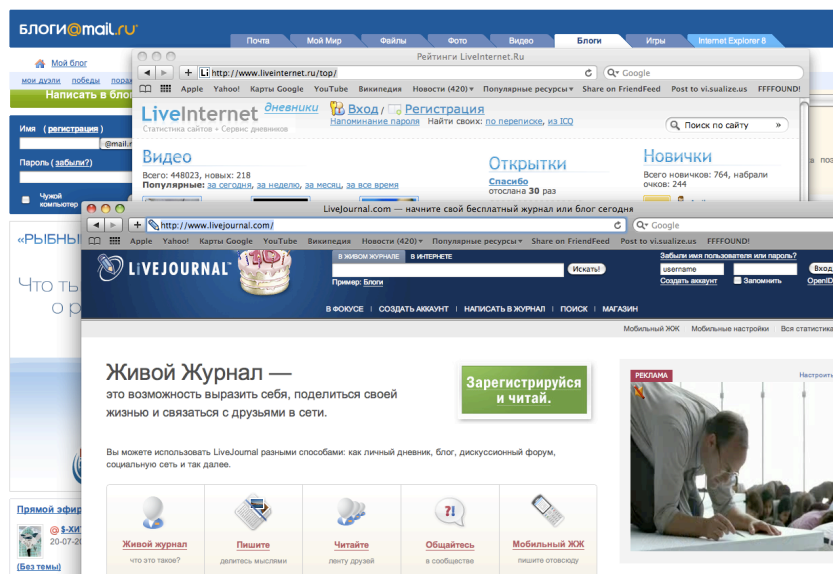
1. Поиск

На какие вопросы отвечает

Чем отличается от поиска по всему интернету

Сервис основан на распространённых в интернете открытых форматах.

Благодаря сотрудничеству с владельцами блог-хостингов эти форматы (RSS, FOAF, Weblogs.Ping) стали стандартом в российской блогосфере.



- Livejournal.com
- Blogs.mail.ru
- Liveinternet.ru

> 190 блог-хостингов

Что индексирует поиск по блогам

- блоги (RSS и ATOM)
- форумы (RSS и ATOM)
- профили (FOAF + Yandex FOAF extension)
- комментарии (RSS и ATOM)

С помощью пингов можно ускорить индексацию, сделав её почти мгновенной

Как происходит индексирование

- На данный момент новые записи индексируются в течение 5 минут с момента их появления на более чем **170 блогхостингах**, включая:
 - LiveJournal.com
 - LiveInternet.ru
 - Blogs.mail.ru
 - Diary.ru
- Индексируются комментарии на LiveJournal.ru, LiveInternet.ru, Blogs.mail.ru и многих автономных блогах
- Проиндексировано более **50 миллионов** профилей, включая профили пользователей пяти крупнейших блог-хостингов

Отличия от веб-поиска

- Очень быстрая индексация: запись попадает в поиск через 1-5 минут после написания
- Свежесть критична: ранжирование по времени
- Много небольших текстов
- Знаем информацию об авторстве и социальных связях и структуре блогов
- Данные не переиндексируются каждый раз заново, а накапливаются в архив блогосферы
 - Существует проблема: RSS не позволяет сообщать об удалении записей – скрыть их из индекса можно только по запросу автора в службу поддержки

2. Пульс блогосферы

Лучше один раз увидеть

Что такое “Пuls блогосферы”?

“Пuls блогосферы” - это служба в Поиске по блогам, с помощью которой можно увидеть, как много записей написали о том или ином явлении в разное время

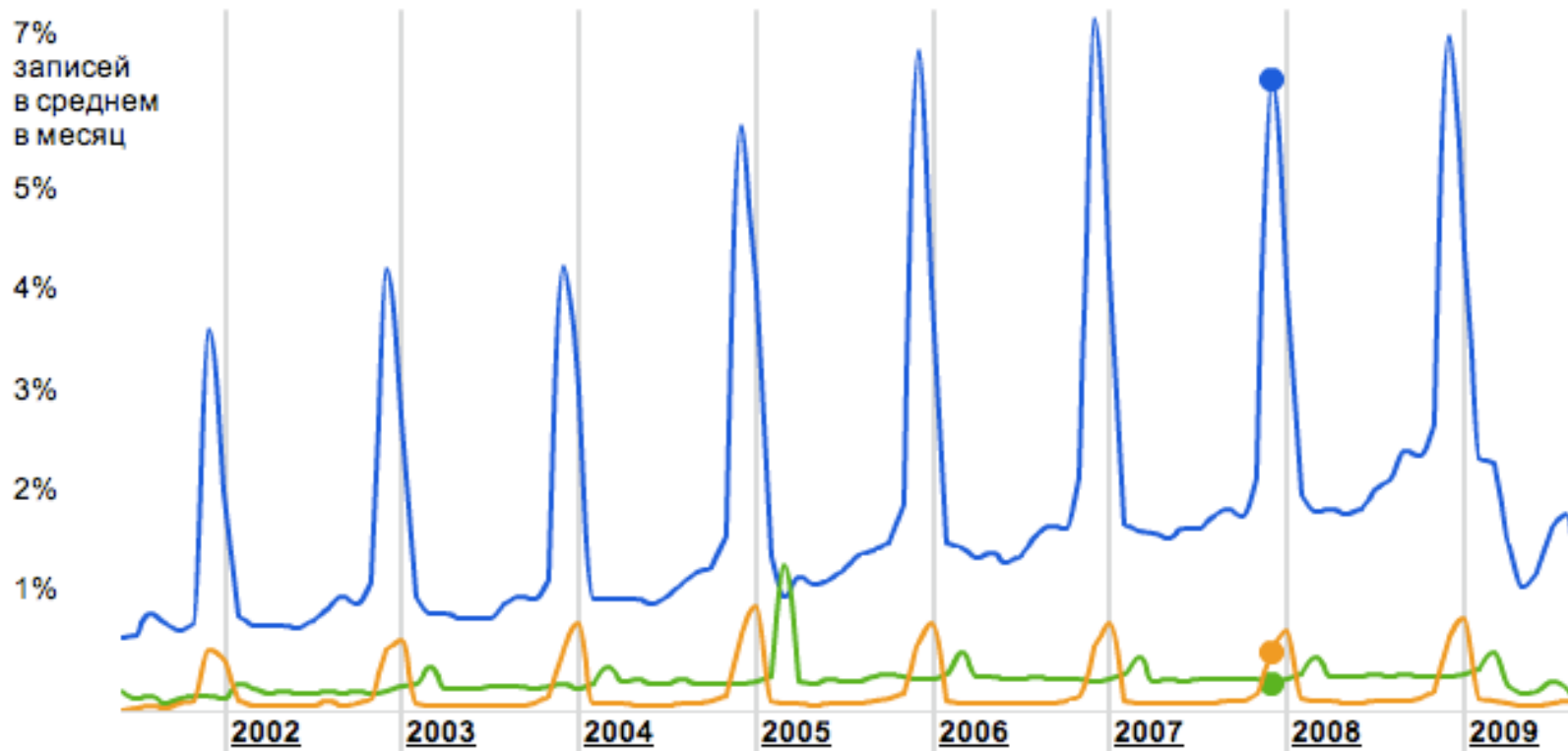
Результаты представлены в виде процентов записей от всех за указанное время

С помощью “Пулса” можно сравнивать обсуждаемость событий в блогосфере, следить за тенденциями в общественном мнении или просто визуализировать популярность явлений

Периодические события

- новый год 6.331%
- женский день 0.269%
- рождество 0.584%

Яндекс

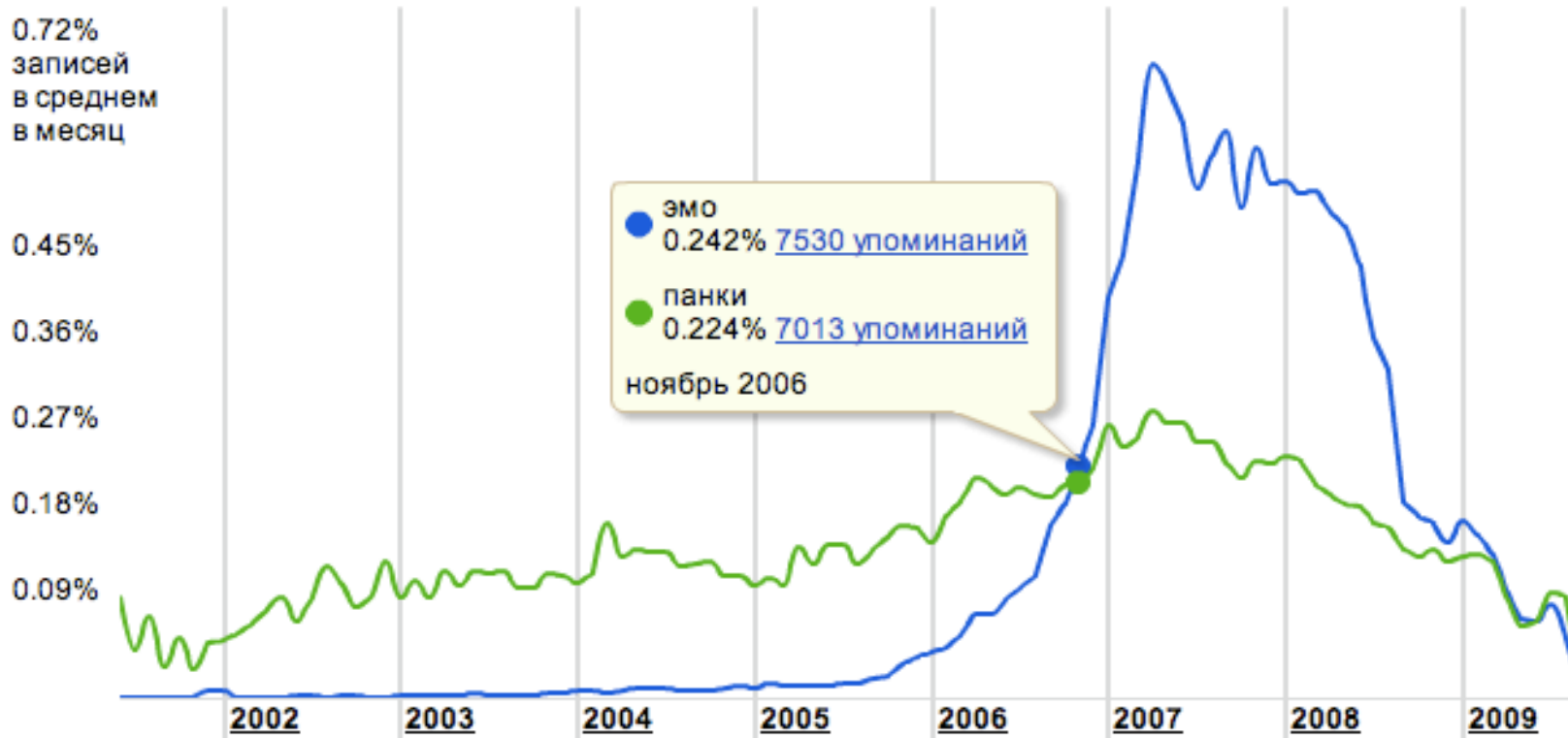


Я

Тенденции

эмо 0.242%
панки 0.224%

Яндекс



Я

3. Рейтинги

Помогают ориентироваться:

Где больше всего пишут

Что обсуждают

Рейтинг блогов

Помогает найти интересные блоги для чтения, узнать какой из блогов сейчас популярнее;

Даёт информацию новичкам о положении дел в блогосфере;

Выделяет самые широко-известные блоги.

Расчитывается на основании данных о ссылках между блогами за последние полгода: чем больше блогов сослалось на разные записи данного, тем он выше в рейтинге.

Рейтинг сервисов

Рейтинг блогхостингов строится ежедневно по количеству записей на каждом сервисе за вчерашний день.

В рейтинге учитывается меньше записей, чем попадает в поиск, не учитываются:

- автоматические записи (например, автопоздравления с днём рождения на Блоги@Mail.ru или “человек опубликовал фото” на Я.ру)
- импортированные записи
- записи автоматических ботов
- спамовые записи

Рейтинги обсуждений

- рейтинуются по количеству упоминаний того или иного объекта;
- рассчитываются за ограниченное время (например, за последние три дня);
- пересчитываются раз в сутки;
- сами рейтингуемые объекты берутся не из блогов, а из готовых источников: например, фильмы из Яндекс.Афиши;
- проблема: пока невозможно автоматически отличать полноценный отзыв от упоминания мимоходом, а также отличать положительные упоминания от отрицательных.

4. Темы дня

Темы дня: "О чём сейчас **многие** говорят?"

Что такое темы дня?

События или явления, больше всего заинтересовавшие блоггеров **сегодня** по сравнению с обычным интересом к ним.

Что больше всего обсуждают сегодня блоггеры.
В противоположность новостям, где событием считается то, о чём больше всего пишут СМИ.

Яндекс

http://www.yandex.by/ RSS ↻ Яндекс

Сделать Яндекс стартовой страницей Настройка ▾ Белорусские мотивы Почта AntonMe Выход

Сегодня в новостях 16:25 **все** в мире

1. Лукашенко считает [ошибкой отказ](#) Белоруссии от ядерного оружия
2. На похороны Леха Качиньского [придут Дмитрий Медведев и Барак Обама](#)
3. ЦИК Беларуси разрешил избирателям предоставлять [для голосования вместо паспорта иные документы](#)
4. Бакиев возвращается в [Джалал-Абад после срыва митинга в Оше](#)
5. Белоруссия (U-18) разгромно [уступила Швеции](#)

Яндекс.dmg Вкусно и полезно

Поиск [Карты](#) [Маркет](#) [Новости](#) [Словари](#) [Блоги](#) [Видео](#) [Картинки](#) ещё ▾ [на-беларуску](#)

Например, [лечебная физкультура при беременности](#) расширенный поиск

Назад, в шестидесятые

Почта
antonius@yandex.ru
233 новых письма
Написать письмо
Подписки
445 сообщений

Фотки
Фото дня

В Минске 15 апреля, четверг, 16:25

[Новая карта Минска](#) [Расписания](#)

Авто европейские иномарки

Маркет летние шины 15"

Словари пациент по-английски

Игры **Развлечения**

Спорт **Бизнес**

Отдых **Музыка**

Учеба **Дом**

Компьютеры **Сайты Минска**

Сегодня в блогах

1. [Новый логотип Femex от Артемия Лебедева](#)
2. [В редакции журнала The New Times изымают документы](#)
3. [Apple одобрила браузер Opera Mini для iPhone](#)

Погода +18
ночью +4, завтра +14

Пробки
[Скачать на мобильный](#)

Телепрограмма

16:00 [Партнеры в действии](#) ВТВ
16:00 [Новости](#) Первый
16:15 [Обручальное кольцо](#) ОНТ

Котировки

	сегодня	завтра
USD НБРБ	2960,00 -5,00	2955,00
EUR НБРБ	4031,82 -14,79	4017,03
RUB НБРБ	101,91 +0,23	102,14

Народ Метрика Директ — запустить генератор продаж

Дизайн — Студия Артемия Лебедева и AntonMe

Русская клавиатура yandex.ru Мобильная версия О компании · About · Вакансии · Реклама · Помощь © 1997—2010 «Яндекс»

Главные три темы дня могут видеть каждый день **12 миллионов** посетителей главной страницы, включая **500+ тысяч** в Беларуси

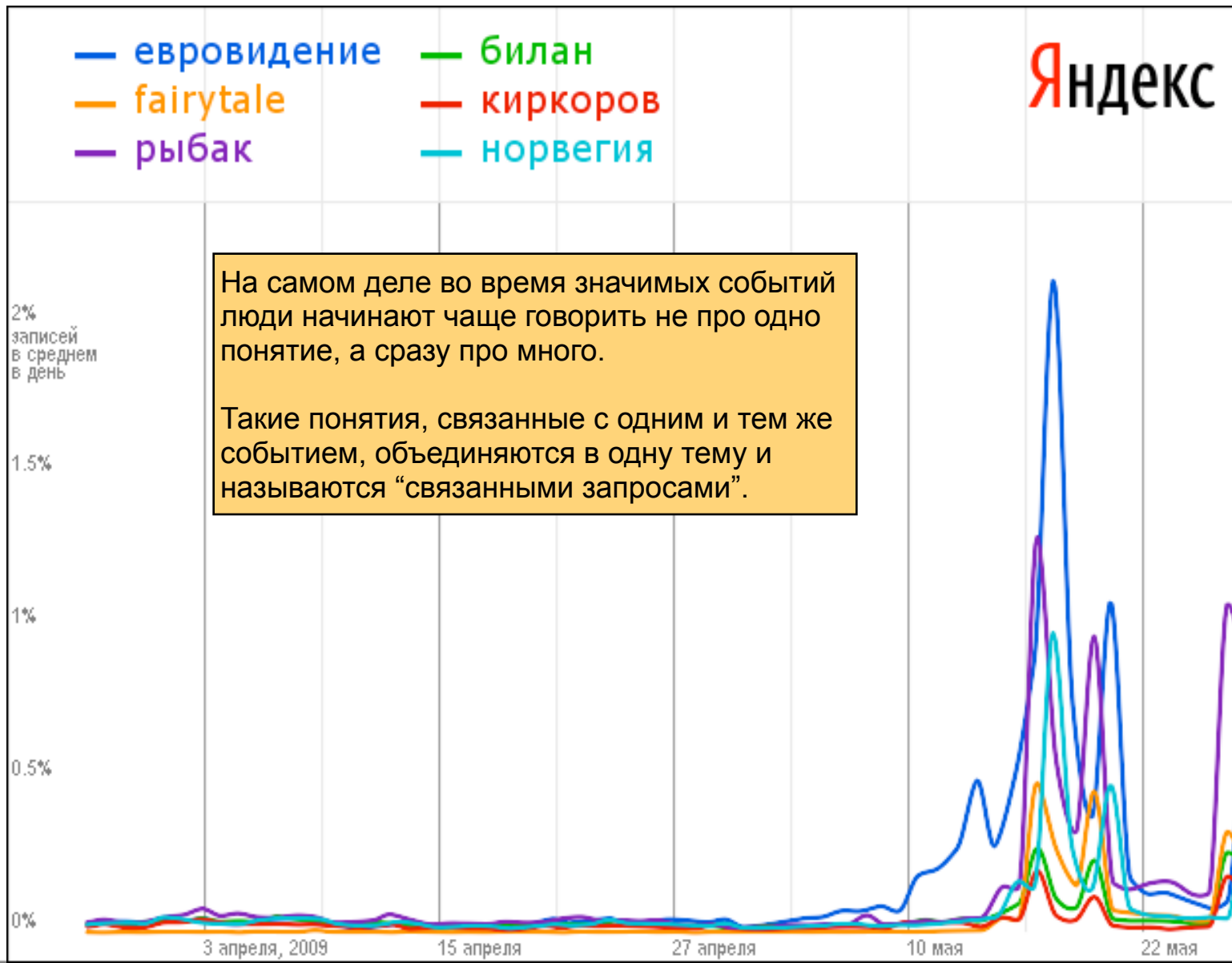
Почему сложно выделять темы дня в блогах?

Новости	Блоги
Пишут о событиях	Пишут и о событиях и о повседневном
Язык, ограниченный жанром и форматом	Свободный, почти разговорный язык
События освещаются похоже	Огромное количество разных способов назвать одно и то же
30 000 новостей в день	300 000 записей в день

Как работают темы дня

- сначала из различных источников выбирается набор гипотез, которые могут оказаться темами
- после этого определяется, как много записей о каждой из них написано сегодня, и как много писали в среднем в прошлом
- те гипотезы, о которых сегодня внезапно стали писать больше записей, чем обычно, считаются темами дня
- близкие темы дня объединяются
- для тем дня выбираются названия
 - проблема: запросы и заголовки записей блоггеров не очень информативны, поэтому параллельно совершенствованию технологии автоматического выбора заголовков, перед обновлением тем дня происходит проверка заголовков модераторами

Близкие темы склеиваются



Как именно склеиваются темы

- Темы дня склеиваются, когда о них часто пишут в одних и тех же записях



5. Открытые данные

API. Какие данные доступны?

- Все свежие проиндексированные записи
- Поиск по всем профилям (FOAF) и данные из них
- Социальные связи (кого “зафрендил” каждый пользователь, кто “зафрендил” его)
- То, что мы смогли понять в результате сбора всех данных в одном месте (агрегирование и анализ):
 - результат определения пола
 - поиск по записям с учётом данных из FOAF
 - фильтрация по блогам, форумам, комментариям и т.п.
- В будущем будут доступны, также, данные из всех рейтингов
- Мы готовы предоставлять для исследований и другие накопленные данные

Вопросы?

ЯНДЕКС

Антон Волнухин

anton@yandex-team.ru